Face Sketch Generation and Recognition

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

Kushal Kumar Jain
2019111001
kushal.kumar@research.iiit.ac.in

Advisor: Dr. Anoop M. Namboodiri

International Institute of Information Technology Hyderabad
500 032, India

May 2025

Copyright © Kushal Kumar Jain, 2025
All Rights Reserved

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis proposal titled ***Face Sketch Generation and Recognition*** by *Kushal Kumar Jain* has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Advisor: Dr. Anoop M. Namboodiri

To all my loved ones.

Acknowledgments

I extend my heartfelt gratitude to my advisor, Dr. Anoop M. Namboodiri, for granting me the opportunity to work at the Center for Visual Information Technology and for his invaluable and truly ineffable guidance throughout my research journey. I am deeply appreciative of Professor Anil K. Jain for his insights, which opened new research directions and enriched my understanding, particularly in applying modern generative models to complex real world problems like crime investigation. I also thank Dr. Manish Srivastava, Dr. Avinash Sharma, and the other professors at IIIT Hyderabad, whose courses sparked my curiosity and laid the foundation for my academic pursuits.

I extend my sincere gratitude to my seniors, Steve Grosz, Sai Amrit Patnaik, and Gowri Lekshmy, for their invaluable guidance and encouragement during various stages of my research journey. I am deeply thankful to my colleagues Ankith Varun, Rahul Goel and Ishaan Shah for being there whenever I needed help. I would also like to thank Abhijit Manatkar, Tejasvi Chebrolu, and Tathagato Roy, for their collaborative spirit and their unwavering confidence in me. Their late-night efforts and insightful discussions have been pivotal to our shared accomplishments.

To my friends and peers at IIIT Hyderabad, your companionship has made this journey unforgettable. You have helped me navigate through extremely rough patches and celebrate successes that have shaped me as a person. Lastly, I express heartfelt appreciation to my sister and parents for their unconditional support and encouragement, without which this journey would not have been possible.

Abstract

The field of sketch generation and recognition has seen significant advancements through the innovative application of generative models. This thesis presents a comprehensive exploration of face stylization , artistic portrait generation and forensic sketch synthesis, leveraging the power of state-of-the-art generative models like StyleGAN and StableDiffusion. Our work addresses key challenges in preserving identity, accommodating various poses, and bridging the modality gap between sketches and photographs. Through three interconnected studies, we demonstrate significant advancements in generating high-quality sketches and improving forensic applications.

We begin by introducing a novel approach to face cartoonization that preserves identity and accommodates various poses. Unlike conditional-GAN methods, our technique utilizes an encoder to capture pose and identity information, generating embeddings within StyleGAN's latent space. This approach uniquely adapts a pre-trained StyleGAN model, originally designed for realistic facial images, to produce cartoonized outputs without requiring a dedicated fine-tuned model.

Building upon this foundation, we present Portrait Sketching StyleGAN (PS-StyleGAN), a style transfer approach tailored for portrait sketch synthesis. PS-StyleGAN leverages StyleGAN's semantic W+ latent space to generate portrait sketches while allowing meaningful edits such as pose and expression alterations. By introducing Attentive Affine transform blocks and a specialized training strategy, we achieve high-quality sketch generation without fine-tuning StyleGAN itself. This method demonstrates superior performance over current state-of-the-art techniques, requiring only a small number of paired examples and minimal training time.

Finally, we address the challenging task of forensic sketch-to-mugshot matching with CLIP4Sketch, a novel approach that uses diffusion models to generate diverse sketch images. By combining CLIP and Adaface embeddings of reference mugshots with textual-style descriptions, we create a comprehensive dataset of sketches corresponding to mugshots. This synthetic data significantly improves the accuracy of sketch-to-mugshot matching in face recognition systems, outperforming training on limited real face sketch data and datasets made by GAN-based methods.

Collectively, these contributions push the boundaries of sketch generation and recognition, offering promising applications in both the artistic and forensic domains.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Humanity's ability to capture experiences through simple sketches dates back millennia. Cave paintings in Lascaux, France [1] serve as early evidence of this innate drive to communicate visually. These primitive depictions, primarily featuring animals and hunting scenes, laid the foundation for more sophisticated forms of artistic expression that would emerge as civilization developed. Around 5000 years ago, we saw the first attempts at capturing individual identity through portraiture, with Egyptian pharaohs being among the first people to be immortalized in stone. Early Egyptian and Persian art prioritized symbolism over realism, portraying subjects based on what they represented rather than how they appeared. In contrast, the Greeks shifted toward depicting what they saw, introducing a realistic yet idealized portrayal of the human form.

The realistic proportions and beauty of Greek sculptures were adopted by Roman artists, who often copied Greek works before creating their own. Romans even began using portraiture as a means to establish and expand their influence. Sketching techniques for capturing a person's likeness with more accuracy further advanced during the Renaissance, a period marked by the rediscovery of classical knowledge and innovation. This revolutionized the way artists approached the human figure, driven by figures such as Leonardo da Vinci, whose meticulous anatomical studies set new standards for precision and realism [2]. Da Vinci's work went beyond capturing likeness; it sought to understand the human body scientifically, bringing unprecedented precision to artistic sketches. His efforts inspired artists to depict the human form with anatomical accuracy and laid the foundation for modern scientific illustration.

The success of artistic expression, particularly in sketches and portraits, is highly dependent on human perception. Our brains possess specialized neural pathways dedicated to facial recognition, making us extraordinarily adept at identifying and remembering faces. This neurological predisposition explains our tendency to see faces in random patterns and our ability to recognize individuals from even simple sketches or linearts [3]. Interestingly, while artists throughout history have successfully captured human likeness, the precise reasons why certain lines and features prove more effective than others in conveying identity are still not very clear [4, 5].

Figure 1.1 (a) the Laocoön and His Sons, a remarkable Hellenistic sculpture from around 200 BCE, showcasing intense emotional drama and anatomical mastery. (b) displays Leonardo da Vinci's meticulous anatomical studies from his notebooks, demonstrating his scientific approach to understanding human form and movement. (c) Le Bouchon by Édouard Manet (1878) represents a shift to modernist sensibilities with its loose, expressive brush and ink technique, characteristic of Manet's spontaneous sketching style. (d) The Plumb-pudding in danger (1805) by James Gillray stands as a masterpiece of British political satire, depicting William Pitt and Napoleon Bonaparte carving up a plum pudding representing the globe, which has earned its place in British history.

Automating this laborious task using computer algorithms for practical applications has been extensively researched. While it is easy for artists to make intuitive decisions about which elements to emphasize, particularly evident in caricature, where exaggerated features can paradoxically create more recognizable sketches, it is precisely what makes synthetic sketch generation particularly challenging. Achieving the balance of abstracting out non-crucial elements while enhancing recognizable features in faces is difficult for algorithms. This complex relationship between detail and recognizability continues to drive innovation in both traditional and digital mediums.

Recent advances in computer vision have opened new frontiers in automated sketch generation and recognition. These developments not only promise to improve our understanding of human visual perception, but also offer practical applications in fields such as law enforcement, criminal investigations, and digital entertainment. However, the challenge of replicating human expertise in sketch creation remains significant, as it requires not just our technical skill, but also an intuitive understanding of which features best capture a subject's essential characteristics.

This thesis explores the challenges of generating realistic faces and cartoons using modern computational approaches, with a particular focus on developing more efficient and accurate methods for automated sketch generation. In the next section I will briefly introduce the generative models that have revolutionized synthetic image generation.

1.1 Evolution of Generative Models

In the last decade, Generative Artificial Intelligence has transformed the field of art, especially after the introduction of Stable Diffusion [6]. Generative models like Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPM) have tremendous potential in high resolution image synthesis, allowing computers to create realistic images and stylized sketches with minimal human intervention. These AI models train on massive datasets, learning to generate realistic portraits, stylized images, and caricatures that capture identity with surprising accuracy. Generative models can be broadly divided into three categories: Deep Learning based, GAN based and Diffusion based.

1.1.1 Deep Learning

Deep learning has revolutionized image processing. Among the most powerful deep learning architectures for image processing tasks are Convolutional Neural Networks (CNNs) [7]. CNNs consist of layers of computational units that extract features from an input image, progressively forming more abstract representations. Each layer produces feature maps—filtered transformations of the input that highlight various aspects such as edges, textures, and shapes. This capability makes CNNs particularly effective for applications such as object detection, recognition, and image generation. One innovative application that leverages these properties is neural style transfer, a technique introduced by Gatys et al. [8], which combines the content of one image with the style of another.

The method proposed by Gatys et al. formalizes style transfer using a total loss function defined as:

$$L_{\text{total}}(\vec{p}, \vec{a}, \vec{x}) = \alpha L_{\text{content}}(\vec{p}, \vec{x}) + \beta L_{\text{style}}(\vec{a}, \vec{x}),$$

where $\vec{p}$ is the content image, $\vec{a}$ is the style image, and $\vec{x}$ is the generated image. The content loss, L_{content}, ensures that the generated image retains the structural and semantic information of the content image, while the style loss, L_{style}, encourages the generated image to adopt the stylistic properties of the style image. Gatys et al. modeled style using the *Gram matrix*, which captures spatial correlations between feature maps in the CNN. This innovative use of the Gram matrix allowed the network to abstractly represent artistic styles, enabling the transfer of textures and patterns across images.

While effective, these approaches have limitations, like the reliance on a single style image often restricts its generalization to diverse styles. Moreover, representing style as uniform texture across the entire image can overlook localized stylistic nuances, such as brush strokes or spatial composition, present in complex artworks. Another challenge lies in balancing content and style: as the style transfer process prioritizes stylistic fidelity, it sometimes distorts the content's structural identity. Despite these drawbacks, this method laid the groundwork for subsequent advancements, such as models capable of handling multiple styles, adaptive style transfer, and improved preservation of content features, broadening the creative and practical applications of style transfer techniques [9–11].

1.1.2 Generative Adversarial Networks

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [12], represent a breakthrough in generative modeling through an adversarial training framework. GANs consist of two neural networks—a generator and a discriminator—that compete against each other in a minimax game. The generator creates synthetic data samples, while the discriminator learns to distinguish between real and generated samples, continuously improving each other's capabilities through this adversarial process.

The fundamental GAN objective function is formalized as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(z|x)] + \mathbb{E}_{z \sim p_{data}(z)}[\log(1 - D(G(z|x)))]$$

where G represents the generator network, D represents the discriminator network, x is the input image, and z is random noise. The generator aims to minimize this objective while the discriminator tries to maximize it, leading to an equilibrium where generated samples become indistinguishable from real data.

Conditional GANs (cGANs) [13] extend this framework by incorporating conditional information into both networks. This conditioning enables more controlled generation, where outputs can be influenced by specific input criteria such as class labels or input images. The conditioning mechanism has proven particularly valuable for tasks like image-to-image translation and style transfer.

StyleGAN [14, 15] further advanced GAN architecture by introducing an intermediate latent space W, which enables fine-grained control over generated images. This innovation allows for semantic manipulation of specific features while maintaining overall image coherence. However, GANs face several challenges, including training instability, mode collapse (where the generator produces limited varieties of outputs), and the difficulty of achieving consistent high-quality results across diverse domains.

Recent developments in GAN architectures have addressed these limitations through techniques such as progressive growing [14], use of adaptive instance normalization, and improved training strategies. These advances have significantly enhanced the quality and controllability of generated images, making GANs particularly effective for applications like face manipulation, style transfer, and sketch generation.

1.1.3 Diffusion Models

While GANs have their advantages, they suffer significant challenges including mode collapse and training instability. Alternative approaches, particularly diffusion models [16–18], have emerged as a powerful solution for high-quality image generation. Text-to-Image models have gained much popularity due to their ease of use [6, 19, 20].

Text-to-Image (T2I) diffusion models are composed of two primary components: the diffusion backbone (ϵ_θ) and the text encoder (c_θ). The diffusion backbone executes a progressive denoising process, beginning with a sample drawn from a noise distribution and iteratively refining it to produce a final denoised image, x_0. This process unfolds across a series of timesteps, where each intermediate state x_t represents a combination of signal and noise, with the signal-to-noise ratio governed by the timestep t. Concurrently, the text encoder processes the input text y through tokenization and embedding lookup, generating initial text embeddings F_{txt}. These embeddings are subsequently refined by the encoder to produce contextualized embeddings E_{txt}, which encapsulate the semantic content and nuances of the textual description.

The integration of textual information into the image generation process is typically achieved through cross-attention mechanisms, which allow the diffusion backbone to leverage the semantic content encoded in E_{txt}. This integration enables the model to generate images that align with the provided textual descriptions. The objective function for T2I diffusion models is formulated as:

$$\mathcal{L} = \mathbb{E}_{x,y,\epsilon,t}\left[||\epsilon\theta\left(x_t, t, c_\theta\left(F_{txt}\right)\right) - \epsilon||_2^2\right] \tag{1.1}$$

where ϵ represents the target noise, and the model aims to minimize the discrepancy between the predicted and actual noise at each denoising step, conditioned on the refined text embeddings. In the next chapter I will briefly cover existing sketch generation and recognition approaches using these models.

Chapter 2

Related Works

2.1 Traditional Approaches for Face-to-Sketch Matching

Early approaches to face-to-sketch generation were predominantly based on hand-crafted features and classical computer vision techniques, encompassing both discriminative and dimensionality reduction methodologies. The foundational work by Tang and Wang [21, 22] introduced exemplar-based face sketch synthesis, utilizing linear transformations between face photos and sketches through Markov random field models. This approach was subsequently enhanced through local linear embedding (LLE) [23], which provided improved handling of non-linear relationships in the transformation process. Many Non-Photorealistic Rendering methods tried to solve the artistic challenge of face sketch synthesis [24–28]. However, they rely on ground truth geometry, which is noisy near detailed parts of the face like eyes, nose and lips.

The field of sketch recognition progressed through various mathematical frameworks, including Partial Linear Squares (PLS) [29] and Coupled Information-Theoretic Projection (CITP) [30], which aimed to bridge the domain gap between photographs and sketches. Local Feature-based Discriminant Analysis (LFDA) [31] advanced the state-of-the-art by incorporating local geometric structures in the feature space. Component-based representation techniques [32] emerged alongside descriptor-based methods, such as the Self-Similarity Descriptor (SSD) dictionary [33], which attempted to capture invariant features across modalities. More recently researchers have used deep learning for feature extraction which has also improved results [34–38], and spawned hybrid approaches like [39], and [40]. In [41], the authors exploit Gaussian Process regression to deal with both the memory gap and modality gap.

These traditional discriminative approaches, while establishing crucial theoretical foundations, were ultimately constrained by their reliance on hand-engineered features and struggled to capture the full complexity of the face-to-sketch translation task, particularly in handling variations in style and preserving fine details. They were also constrained by small datasets and limited compute of the time.

2.2 Deep Learning and GAN based Approaches

The advent of Generative Adversarial Networks (GANs) [12] and conditional GANs (cGANs) [13] represented a paradigm shift in face-to-sketch generation. The introduction of frameworks such as Pix2Pix [42] established fundamental architectures for image-to-image translation. Recent methods, such as [43], have incorporated Dynamic Normalization (DySPADE) in the generator architecture, alongside depth maps for supervision, yielding promising results in sketch generation tasks. The introduction of unsupervised models like CycleGAN [44], along with advancements like UNIT and UGATIT [45, 46], has led to the development of numerous compelling sketch generation techniques.

In particular, Semi-Cycle-GAN (SCG) [47] proposed a semi-supervised approach with a noise-injection strategy to overcome the challenge of missing features in CycleGAN outputs. AP-DrawingGAN [48] utilized specialized GANs to handle challenging facial features such as eyes, nose, and lips. Extending this, FSGAN [49] introduced the FS2K dataset, which comprises three different styles with paired sketch examples. However, the efficacy of these approaches in face recognition hinges significantly on the quality of their synthesis results. More recently, methods like [50] have leveraged CLIP-based losses [51] to enhance sketch generation. Despite these successes, these approaches often require extensive training of the generator, which can be computationally intensive and demands large datasets. This requirement poses a significant challenge, particularly for tasks like face sketching, where such large datasets are not readily available.

2.3 StyleGAN based Approaches

StyleGAN emerged as a particularly transformative architecture, distinguished by its exceptional image quality and rich semantic representation capabilities. The architecture's success led to numerous innovations, particularly in the development of the $W+$ space [52], which significantly enhanced reconstruction capabilities while maintaining precise editing control. The latent spaces of GANs have been the focus of significant research, particularly with regard to exploring and controlling their semantic properties. One emerging task in this area is GAN inversion, which seeks to map a given image back to its corresponding latent vector in a pre-trained GAN, enabling reconstruction of the image from its latent representation [53]. StyleGAN [15] has become a preferred model for such investigations, owing to its highly expressive latent space, which captures a wide range of semantic features in a compact and disentangled manner.

There are two primary approaches to GAN inversion. The first involves directly optimizing the latent vector to minimize the reconstruction error for each individual image [52, 54]. The second approach, inspired by Richardson et al. (2020) [55], trains an encoder to map images to their corresponding latent vectors, bypassing the need for iterative optimization. Our method follows the latter approach, where we design an encoder that directly maps input conditions onto the StyleGAN latent space, avoiding the "invert first, edit later" paradigm commonly used in other frameworks [56].

The evolution of StyleGAN-based approaches has been marked by several significant architectural advances that allowed stylization with editability. DualStyleGAN [57], JoJoGAN [58] have demonstrated remarkable success in style transfer while maintaining semantic editing capabilities. More recent advancements, like StyleTransformer [59] and TransStyleGAN [60] incorporate attention mechanisms to enhance the inversion tasks. These improvements are crucial when transforming style codes in the later stages of generation, especially for tasks like portrait sketches, where spatial consistency is essential for maintaining well-defined structures.

2.4 Diffusion Based Approches

The introduction of diffusion models marks the latest significant advancement in face-to-sketch generation, offering a fundamentally different approach to image synthesis. These models operate through an iterative denoising process, providing several inherent advantages over their GAN-based predecessors like a more tractable generation procedure that can be conditioned with text embeddings or identity embeddings. Various strategies have emerged to preserve and personalize face identity of generated images of diffusion models. While some techniques, such as Textual Inversion [61] and DreamBooth [62], necessitate fine-tuning for each new concept, others like IP-Adapter [63] and PhotoMaker [64] offer the advantage of producing identity-consistent generations for multiple subjects without the need for inference time fine-tuning. IP-Adapter and PhotoMaker distinguish themselves by embedding the identity of input reference images into the diffusion process through cross-attention layers.

Chapter 3

Learning Latent StyleGAN Representations for Stylization

Among all the artistic transformations enabled by synthetic image generation, in this chapter we focus on face cartoonization but we will use the same setup in the next chapter where we focus on sketch generation. Cartoonization typically involves abstracting real-world images by simplifying shapes and applying smooth color shading, resulting in images with distinct visual qualities such as well-defined edges and minimal texture [65]. Sketch generation, on the other hand, aims to produce line-based representations of faces, often used in artistic and forensic applications. Both these tasks present unique challenges in preserving identity, accommodating various poses, and bridging the modality gap between stylized representations and photographs.

3.1 Method

Our approach generates a cartoon image representing the identity of one input image (I_d) and other attributes, primarily pose and expression, of another input image (I_p). The model architecture comprises encoders, a Multi-Layer Perceptron (MLP), and generators. We employ a pre-trained ArcFace model for identity encoding (E_{id}) and a fine-tuned VGG model for pose encoding (E_p). These embeddings are concatenated and passed through the MLP to produce a vector w, which serves as input to the generator:

$$w = M([E_p(I_p), E_{id}(I_{id})]) \tag{3.1}$$

The generator consists of two components: a pre-trained StyleGAN (G) and a GAN-based cartoon generator called White Box Cartoonizer (C). The StyleGAN, pre-trained on the FFHQ dataset, acts as a pre-cartoonization backbone. The White Box Cartoonizer, building upon CartoonGAN, breaks down the cartoonization process into surface, structure, and texture representation, offering more control and tunability. The output image is generated as follows:

$$I_{out} = C(G(w)) \tag{3.2}$$

To ensure the quality and accuracy of the generated cartoon images, we employ multiple loss functions. The identity loss ($\mathcal{L}_{id}$) maintains the essential identity attributes in the cartoon space. A landmark

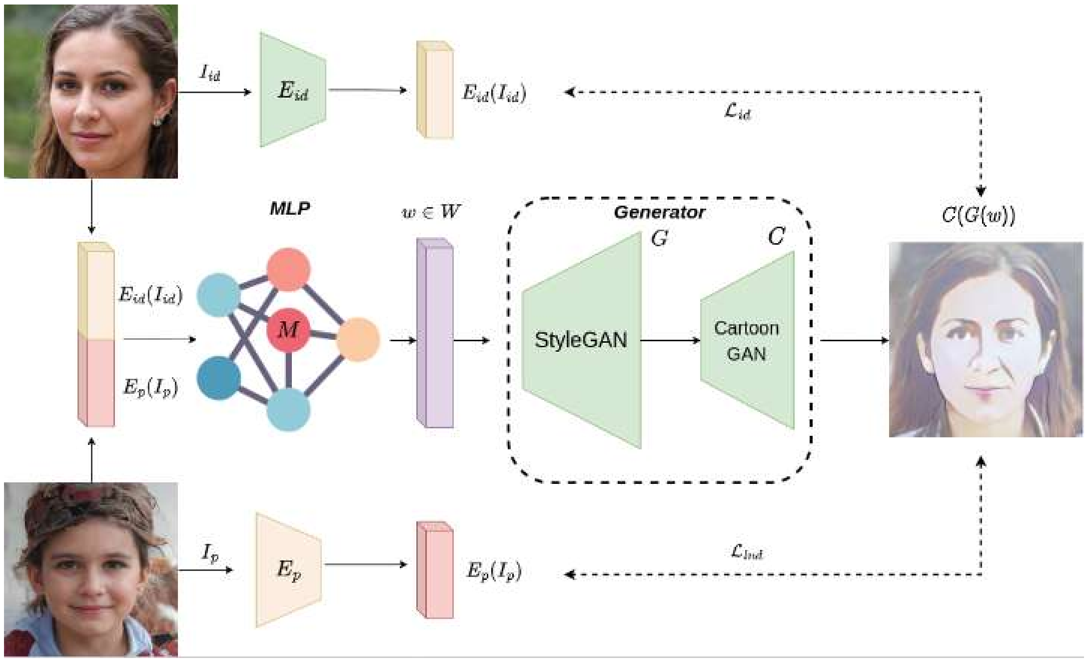

Figure 3.1 Our model architecture. All the models in green are pretrained, while the pose encoder in yellow is being fine-tuned and the MLP is trained from scratch. Data flow is marked with solid lines and losses are marked with dashed lines. Bold-dashed lines enclose the generator setup. Input images I_{id} and I_p are encoded using E_{id} and E_p respectively, to give embeddings. The embeddings are concatenated and passed through the MLP which maps it to the W latent space of StyleGAN. The w vector generated is passed through our generator setup to give the final output $C(G(w))$. $\mathcal{L}_{id}$ ensures that identity is preserved, $\mathcal{L}_{lnd}$ enforces pose translation. Another loss term $\mathcal{L}_{rec}$ is also used but is not shown here.

loss ($\mathcal{L}_{lnd}$) ensures accurate transfer of pose and facial arrangement. Additionally, a reconstruction loss ($\mathcal{L}_{rec}$) is applied when $I_{id} = I_p$ to preserve non-facial attributes like illumination and color. This reconstruction loss is a weighted sum of L_1 and MS-SSIM losses, applied only when the identity and pose images are identical to prevent unwanted reconstruction of identity features from I_p. $\mathcal{L}_{mix}$ is a weighted sum of L_1 loss and MS-SSIM loss with α being a hyper-parameter.

$$\mathcal{L}_{mix} = \alpha(1 - MS\text{-}SSIM(I_p, I_{out})) + (1-\alpha)\|I_p - I_{out}\|_1 \tag{3.3}$$

We impose a constraint on the application of this loss, limiting it exclusively to cases where the identity and pose images are identical. This deliberate constraint is designed to prevent the reconstruction of identity features from I_p within the resulting cartoon image. The below redefinition of $\mathcal{L}_{mix}$ takes this into account.

$$\mathcal{L}_{rec} = \begin{cases} \mathcal{L}_{mix} & \text{if } I_{id} = I_p \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

These losses collectively guide the model to generate cartoon images that balance stylistic transformation with faithful representation of identity and pose.

Implementation Details: We utilize a StyleGAN pre-trained at 256x256 resolution. The ratio of training samples with $I_{id} = I_p$ and $I_{id} \neq I_p$ is a crucial hyper-parameter that controls the weight for disentanglement and reconstruction. Following the suggestion in [66], we use $I_{id} = I_p$ every third iteration, and $I_{id} \neq I_p$ otherwise. The landmarks regression network E_{lnd} is implemented using a pre-trained model [67], trained to regress 68 facial keypoints. For optimizing $\mathcal{L}_{total}$, we use a learning rate of $5e^{-5}$. The loss weights are set as $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.001$, $\alpha = 0.84$.

3.2 Interpolating Between Poses and Identities

Figure 3.2 Results of pose/expression (first 3 rows) and identity (last 3 rows) interpolation. We sample eight vectors in the linearly interpolated space between the mapped latent codes w_1 and w_2 using $w = w_1 + (w_2 - w_1)\frac{k}{8}$ where $k \in \{1, 2, ..., 8\}$, and pass it through our generator $C(G(w))$ to obtain the cartoonized images.

Previous works have demonstrated that the W space exhibits a high degree of controllability, allowing for seamless feature manipulation through latent code interpolation [68, 69]. Our method identifies latent codes that are optimal for cartoonization while simultaneously capturing the high-dimensional identity attribute and an intricate combination of expression and pose.

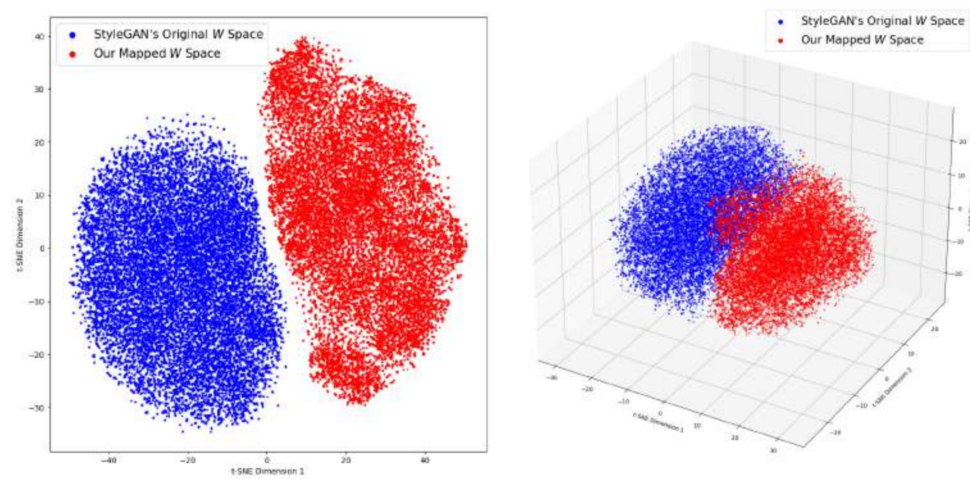

Figure 3.3 t-SNE visualization of the latent vectors obtained from the two distributions : StyleGAN's original W space and our mapped W space

Figure 3.2 visually demonstrates the disentanglement of these high-level features through latent code interpolation, showcasing the robust nature of StyleGAN's latent space W and its resilience to style distortion. The interpolation experiments are conducted as follows:

- **Pose and Expression Interpolation:** In the first two rows of Figure 3.2, we keep the identity fixed and extract the attribute features (representing pose and expression) from two different images. We then obtain two corresponding latent codes w_1 and w_2 using our proposed method and linearly interpolate between them.

- **Identity Interpolation:** For the last two rows, we keep the attribute features (pose and expression) fixed and interpolate between two different identities.

3.3 The New Learned Space

To validate our hypothesis that the encoder successfully learned a novel distribution within the StyleGAN latent space, we decided to use t-SNE [70] to visualise the vectors for comparison. Specifically, we aimed to compare the original w vectors from StyleGAN with the vectors produced by our trained MLP. The MLP-generated vectors are optimized for maintaining identity during the cartoonization process, as opposed to generating photorealistic images. As illustrated in Figure 3.3, the two distinct clusters observed—one representing the MLP-generated vectors and the other the original StyleGAN w vectors—confirm the separability of these two distributions.

3.4 Chapter Conclusion

In this chapter, we introduced a novel approach to face cartoonization leveraging the power of StyleGAN's latent space. Our method successfully disentangles identity and attribute features, allowing for precise control over the cartoonization process. By training a custom encoder and utilizing a combination of pre-trained models, we achieved high-quality cartoon outputs that maintain identity while accurately transferring pose and expression. The effectiveness of our approach was demonstrated through various experiments, including latent space interpolation and t-SNE visualization, which confirmed the creation of a new, optimized distribution within the StyleGAN latent space.

While our method shows promising results in the realm of face cartoonization, it relies on fine-tuning and training specific components for the task at hand. This approach, although effective, may not be optimal as it changes the initial W latent space, hence rendering pretrained encoders for StyleGAN useless. Adapting outputs of StyleGAN without changing the initial W space is the key challenge that leads us to our next chapter, ”PS-StyleGAN: Making Sketches Without Fine-tuning,” where we explore a more flexible and efficient approach to sketch generation that doesn't require extensive model adjustments or training.

Chapter 4

PS-StyleGAN : Portrait Sketching Without Finetuning

While cartoonization involves a transformation to a more abstract domain, sketch generation presents its own set of challenges. Traditional portrait sketches rely on minimalistic lines and shading to capture both the identity and expressive features of a subject [4, 5]. However, synthetic generation of such sketches can be difficult, especially when it comes to preserving the fine details that are critical to facial recognition, such as the eyes and lips. Previous non-photorealistic rendering techniques [24–28] often rely on ground-truth geometry, which can be noisy or imprecise in detailed areas, making deep learning-based approaches more appealing for this task.

To address the specific challenges of sketch generation, we propose Portrait Sketching StyleGAN (PS-StyleGAN), a novel method that leverages the existing capabilities of StyleGAN without the need for extensive fine-tuning. Unlike methods that modify the generator or latent space, PS-StyleGAN focuses on stylistic transformations while maintaining the integrity of the pre-trained StyleGAN model. By introducing attention-based affine transformation blocks, we are able to apply sketch-style transformations directly to the output, ensuring that critical facial features are preserved while producing realistic and artistically expressive sketches. This approach allows us to retain the flexibility and semantic control of StyleGAN's latent space, making it compatible with existing encoders designed for the original StyleGAN framework.

4.1 Method

We propose a novel style transfer approach that utilizes *attentive affine* transformations to improve style adaptation at fine layers of the StyleGAN architecture. This method leverages the disentangled latent space of StyleGAN to generate more expressive style features by computing attention maps between the content and style latent codes. Our approach is divided into two stages: Domain Transfer and Conditional Refinement, which progressively align the generative space to the target style distribution.

4.1.1 Attentive Affine Transformations

We introduce an *attentive affine* transformation to learn better affine parameters at fine layers, which represent the full feature distribution of the style image. To achieve this, we first compute attention

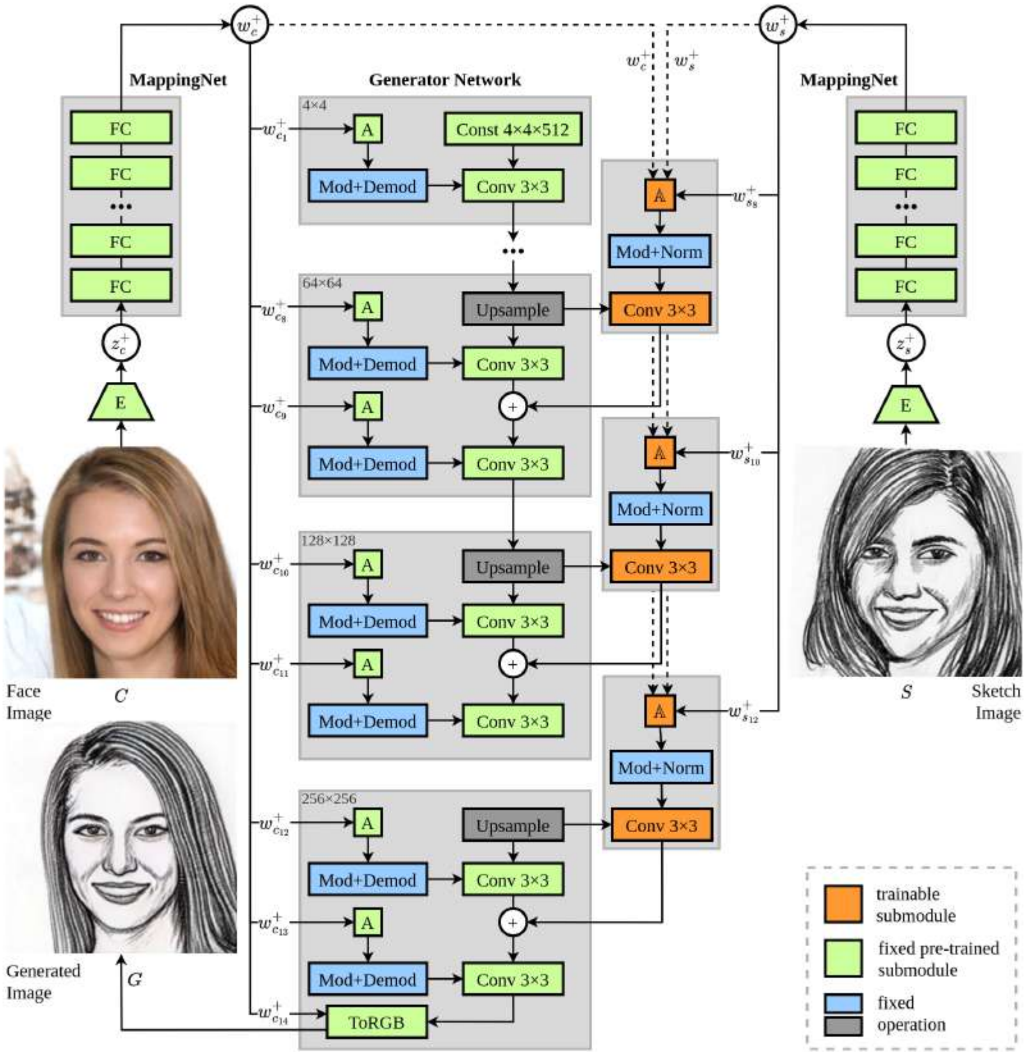

Figure 4.1 An overview of our model architecture. We use a pretrained 256x256 resolution StyleGAN2 [15] generator g fitted with three style adaptation blocks at the fine resolution layers. Each block consists of a novel Attentive Affine transform module ($\mathbb{A}$) that predicts affine parameters from attention-weighted latent codes of S using supervision from w_c^+ and w_s^+. These parameters are then used to modulate and normalize the spatial features of g at different scales to imbibe the style $\mathbb{S}$ into C.

maps using the content and style latent codes, w_c^+ and w_s^+, respectively. The attention map A measures similarity in the StyleGAN's latent space, allowing us to capture relationships between style and content features efficiently. The attention map is computed as:

$$A = \mathit{Softmax}(Q^T \otimes K) \tag{4.1}$$

where Q, K, and V are derived from normalized latent codes through 1x1 convolution layers.

Figure 4.2 Outputs of PS-StyleGAN for different sketching styles (inset) in specified poses and expressions while maintaining the input identity. A model trained on FS2K dataset was used for (b) - (d), while CUHK and APDrawing were used for the models in (e) and (f).

4.1.2 Improved Affine Parameters

To extract meaningful style features, we compute an *attention-weighted* latent code segment, $x_{s_i}^{+}$, by multiplying the attention map with the style latent code. The resulting weighted latent code segment is passed through a fully connected layer to produce the improved affine parameters $y_{s,i}^{S}$ and $y_{b,i}^{S}$, as follows:

$$(y_{s,i}^{S}, y_{b,i}^{S}) = \mathit{Affine}(x_{s_i}^{+}) \tag{4.2}$$

These parameters are used to adaptively normalize the content feature map for instance-wise alignment, effectively improving the quality of the generated images in the fine layers.

4.1.3 Adaptive Normalization

We use the improved affine parameters to modulate the content features via an AdaIN operation:

$$F_i^{CS} = y_{s,i}^{S} \frac{F_i^{C} - \mu(F_i^{C})}{\sigma(F_i^{C})} + y_{b,i}^{S} \tag{4.3}$$

where F_i^C is the normalized content feature map. This operation ensures that the content features are aligned with the style distribution in a way that preserves both content and style details in the generated images.

4.1.4 Training Strategy

Our training follows a progressive transfer learning scheme with two stages. In **Stage I**, we perform domain transfer by fine-tuning a pretrained StyleGAN to align its generative space with the target style domain. We use a combination of style loss, content loss, and adversarial loss to fit the generated image to the target style.

In **Stage II**, we refine the model using paired data of photo-realistic images and their corresponding sketches to broaden the style diversity. We incorporate perceptual loss and regularization to prevent

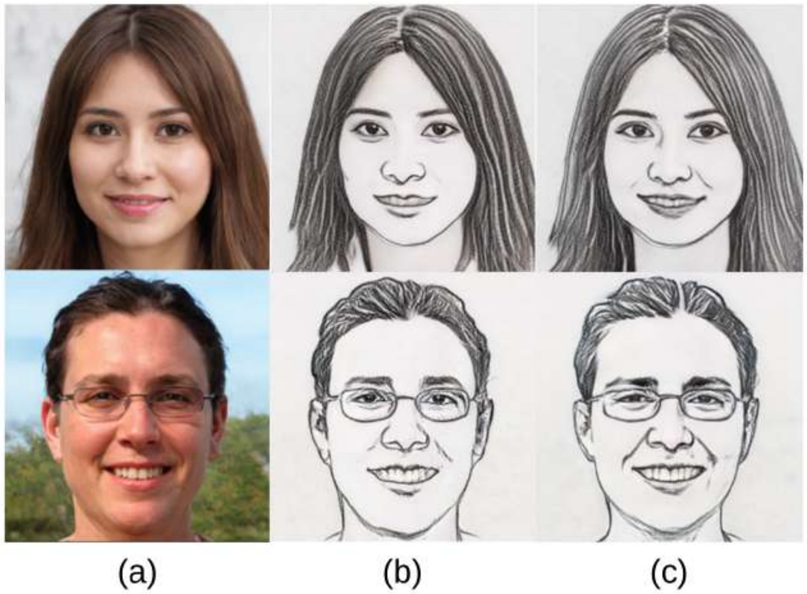

Figure 4.3 Results after each stage of progressive transfer learning. At the end of stage I, the model converges to an average representative style as seen in (b) where the eyes, nose and mouth are sketched in a similar manner. Stage II widens the model's generative space to capture subtle style variations resulting in better identity preservation as shown in (c).

overfitting, adjusting the objective function to further enhance the quality and variability of generated images.

4.2 Qualitative Results

Visually comparing our PS-StyleGAN with leading methods, namely FSGAN [49], HIDA [43], DualStyleGAN [57], and AdaAttN [10], we observe that our method excels in rendering eyes and lips, showcasing sharper details and enhanced realism, see Fig 4.4. Our results are visually most similar to DualStyleGAN but their method also learns shape biases in the dataset hence affecting recognizability. DualStyleGAN often changes the gaze direction and shape of lips too. The Attentive Affine transform blocks in PS-StyleGAN contribute to a superior balance between artistic expression and accuracy, resulting in more visually appealing and faithful representations of facial features.

4.3 Quantitative Results

To objectively assess the performance of our method, we employ four evaluation metrics: Learned Perceptual Image Patch Similarity (LPIPS) [71], Structure Co-Occurrence Texture (SCOOT) [72], Fea-

Figure 4.4 Comparison of our method with other state of the art methods on the 3 styles (inset) of FS2K: style 1 (row 1), style 2 (row 2), style 3 (row 3). From left to right: Input identity image, Ours, DualStyleGAN [57], HIDA [43], FSGAN [49], AdaAttN [10].

ture Similarity Index Measure (FSIM) [73], and ID loss [74]. A lower LPIPS and ID loss value indicates a more accurate and realistic generated sketch, while higher SCOOT and FSIM values reflect better resemblance to artist-rendered sketches. The average values of SCOOT, LPIPS, FSIM, and ID loss for all test samples are reported in Table 4.1.

The quantitative results, as presented in Table 4.1, illustrate PS-StyleGAN's superior performance compared to existing methods across multiple evaluation metrics. PS-StyleGAN achieved the highest Structure Co-Occurrence Texture (SCOOT) score (0.5603) and the lowest Learned Perceptual Image Patch Similarity (LPIPS) score (0.2303), indicating a strong balance between style realism and identity

Method	SCOOT ↑	LPIPS ↓	FSIM ↑	ID ↓
HIDA	0.4433	0.3214	0.3660	0.0241
FSGAN	0.3621	0.2890	0.3692	0.0424
AdaAttN	0.4670	0.2600	0.3806	0.0233
DualStyleGAN	0.4490	0.3012	0.3631	0.0247
Ours	**0.5603**	**0.2303**	**0.4283**	**0.0206**

Table 4.1 Quantitative comparison of AdaAttN [10], FSGAN [49], HIDA [43] and DualStyleGAN [57] with our method based on SCOOT, LPIPS, FSIM and ID loss. Our method shows considerably better SCOOT and ID loss values indicating more visually appealing and recognizable results.

preservation. Additionally, PS-StyleGAN surpassed other methods in Feature Similarity Index Measure (FSIM) with a score of 0.4283, reflecting its ability to produce visually appealing sketches that closely resemble the structure of real sketches. The model also achieved the lowest ID loss (0.0206), underscoring its effectiveness in maintaining recognizable identity features, especially compared to models like FSGAN and DualStyleGAN, which exhibited higher ID loss values. These results underscore PS-StyleGAN's strength in generating sketches that combine stylistic fidelity with a high degree of identity consistency, making it particularly suitable for applications requiring detailed facial recognition.

4.4 Chapter Conclusion

In this chapter we presented PS-StyleGAN, a novel approach to portrait sketch generation that overcomes limitations of fine-tuning in traditional StyleGAN-based methods. By introducing attentive affine transformation blocks, PS-StyleGAN enables stylistic control without altering the foundational StyleGAN model. This method allows for flexible, high-quality sketch generation that accurately preserves key facial features critical for identity while accommodating stylistic variations. The attentive affine transformation mechanism effectively modulates the content and style latent codes, capturing expressive details that enhance both realism and artistic quality, when compared to other leading methods like DualStyleGAN and AdaAttN. PS-StyleGAN's strength lies in its adaptability and computational efficiency since it uses a GAN framework, making it a valuable tool for applications where compute is limited. In the next chapter we will explore CLIP4Sketch, a diffusion based method that generates identity-consistent sketches by introducing text-driven and image-conditioned controls.

Chapter 5

CLIP4Sketch : Diffusion Based Sketch Generation

Facial recognition technology has made significant strides in recent years, with advancements in areas such as ArcFace, AdaFace, and DCFace [75–77]. However, a persistent challenge in this field is the accurate matching of hand-drawn forensic sketches to corresponding mugshot photographs. This task is crucial in various domains, including law enforcement, surveillance, and forensic investigations. The inherent difficulty lies in the substantial discrepancy between these two data modalities: forensic sketches are often incomplete artistic interpretations based on witness descriptions, while mugshots are photographic images captured under controlled conditions. Traditional approaches to this problem have primarily relied on discriminative frameworks, but their efficacy has been limited by the scarcity of annotated forensic sketch data available for training and evaluation.

To address this data scarcity issue, we propose CLIP4Sketch, a novel approach that leverages diffusion models to generate sketch images with large variations, enhancing the performance of face recognition systems in sketch-to-mugshot matching. Diffusion models, known for their ability to generate high-quality samples from complex data distributions, offer a promising solution to bridge the gap between sketches and photographs. Unlike previous approaches that predominantly employed Generative Adversarial Networks (GANs), which are susceptible to issues such as mode collapse and limited generalization, our method utilizes the robust capabilities of diffusion models to create a diverse and representative synthetic dataset.

CLIP4Sketch offers a controllable latent diffusion model that uses text and image embeddings as conditions for highly realistic and diverse synthetic sketch generation. Inspired by techniques like IP-Adapter [63] and similar approaches [64], our pipeline integrates ControlNet for spatial control and combines CLIP with Adaface for identity control. This allows us to guide the generation process with explicit control over both identity and style. By aggregating data from multiple publicly available sketch datasets [22, 78–80], we have created a synthetic dataset containing 27,000 unique identities, each with four hand-drawn style sketches and four software-generated style sketches. This diverse dataset enables us to finetune state-of-the-art face recognition models, significantly improving their performance in sketch-to-mugshot matching tasks and advancing the field of sketch-based facial recognition.

Figure 5.1 Qualitative comparison of our CLIP4Sketch generated sketches and real sketches. The second row is generated by our method, while the first row contains examples from existing datasets PRIP-Composites [79] and CUHK [22]. The text prompts used to generate the images in the second row using the proposed CLIP4Sketch model are, in order from left to right, “a viewed software-generated sketch of a face”, “a software-generated sketch of a face”, “a viewed hand-drawn sketch of a face”, and “a hand-drawn sketch of a face”. The leftmost image is the input identity.

5.1 Method

5.1.1 Face-ID Preservation

Recent advancements in face identity preservation for diffusion models have yielded various strategies, ranging from fine-tuning approaches to more flexible methods. While techniques such as Textual Inversion [61] and DreamBooth [62] require model fine-tuning for each new concept, more recent approaches like IP-Adapter [63] and PhotoMaker [64] offer the advantage of generating identity-consistent images for multiple subjects without the need for inference-time fine-tuning. These latter methods embed the identity of input reference images into the diffusion process through cross-attention layers, guiding the model to generate images that align with the identities portrayed in the reference images. Despite the promise of IP-Adapter, empirical evaluation revealed limitations in its spatial control for preserving facial identity. To address this, we opted for a ControlNet [81] architecture in conjunction with an image prompt, enabling more fine-grained spatial control. ControlNet is designed to introduce conditional control to large pretrained text-to-image diffusion models, leveraging source models like Latent Diffusion Models (LDM) and reusing their deep encoding layers [6, 81]. This approach constructs a robust encoder capable of learning specific conditions such as canny edge maps and depth maps. In our implementation, we utilize a canny edge detector to generate an edge map of the input mugshot, which is then incorporated into the diffusion pipeline via the ControlNet adapter. This method enhances the model’s ability to maintain facial identity while allowing for greater control over the generated output.

5.1.2 CLIP4Sketch Pipeline

The CLIP4Sketch pipeline presents an innovative approach to generating sketch portraits from mugshot images while preserving identity and introducing stylistic variations. This method leverages a combination of pre-trained models and cross-attention mechanisms to achieve fine-grained control over both identity and style attributes in the generated sketches.

The process begins by encoding an input mugshot image X_m using both a pre-trained CLIP model E_{clip} and AdaFace E_{fr}, a robust face recognition model. The resulting embeddings are concatenated to form a combined embedding vector C:

$$C = [E_{\text{clip}}(X_m), E_{\text{fr}}(X_m)] \tag{5.1}$$

This combined embedding captures both high-level semantic information and critical identity features. It is then transformed through a Projection Network P to align with the latent space of textual embeddings.

To introduce stylistic variations, the pipeline incorporates textual captions processed by CLIP's text encoder E_{text}. These textual embeddings are appended to the projected combined embeddings and passed through cross-attention layers in the generator. The diffusion model, responsible for sketch generation, utilizes these final embeddings as conditioning input, along with a Canny edge map serving as the ControlNet input.

The CLIP4Sketch method allows for the creation of diverse sketches from a single mugshot by inputting different textual descriptions, which guide the cross-attention layers to adjust the sketch generation process accordingly. During training, the weights ϕ of the image cross-attention layer and the projection network P are optimized, while the weights ψ of the text cross-attention layer and the combined weights θ of the T2I diffusion model remain frozen.

This approach effectively bridges the modality gap between photographs and sketches, enabling the generation of a large, diverse sketch dataset. Using this pipeline, a synthetic dataset of 245,376 sketches representing 27,264 identities across four distinct styles was created. This dataset plays a crucial role in enhancing face recognition models' ability to generalize across various sketch styles, particularly in forensic sketch-to-mugshot matching tasks.

5.2 Datasets

We use multiple sketch and mugshot datasets to train our sketch generation model, a summary of the datasets used can be found in the table 5.1. The publicly available datasets included the CUHK Face Sketch Database (CUHK) [22], PRIP-Composites [79], FS2K [78], TUFTS Face Database [80], IIIT-D Forensic Sketch Database [82], and WildSketch [83]. The CUHK, PRIP-Composites, and FS2K datasets, comprising a total of 8,912 sketches representing 4,460 unique identities, were used exclusively for training the CLIP4Sketch pipeline. For testing the FR system, we employed the TUFTS Face

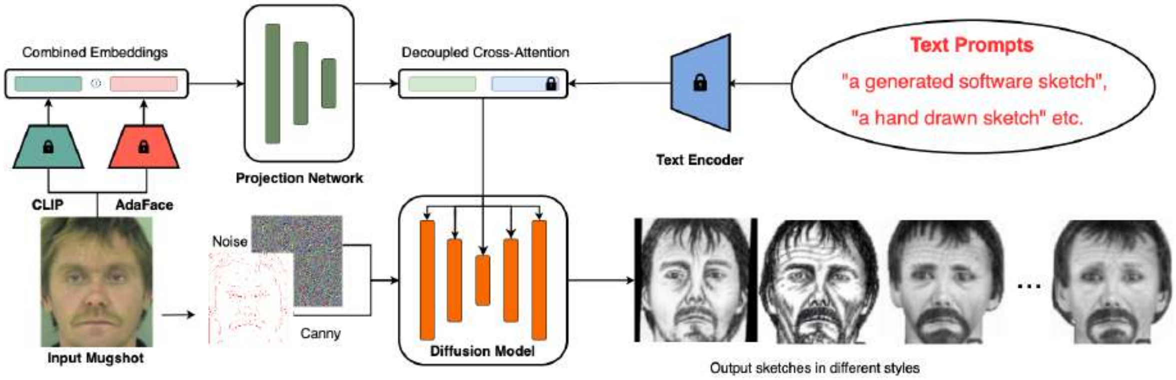

Figure 5.2 The CLIP4Sketch pipeline for generating diverse sketches from a mugshot image. A latent diffusion model is employed which combines embeddings from CLIP and AdaFace to preserve identity and uses text prompts to control stylistic variations. We use a canny edge image as controlnet condition along with decoupled cross-attention layers for identity and style conditioning.

Database, IIIT-D Forensic Sketch Database, and WildSketch, collectively offering 1860 sketches of 980 identities, to provide a realistic evaluation of our model's performance.

5.3 Evaluating Realism of Generated Dataset

To assess the quality of sketches produced by our CLIP4Sketch model, we conducted a rigorous comparative analysis against genuine forensic sketch images. This evaluation focused on two critical dimensions: similarity score distributions and VGG-feature space representations.

5.3.1 Similarity Score Analysis

We employed a pre-trained AdaFace network [76] to extract features from both our generated sketches and real forensic sketches. Using these features, we computed similarity scores for two types of matches:

- **Genuine matches:** Sketches paired with their corresponding mugshots
- **Imposter matches:** Sketches paired with non-corresponding mugshots

Our analysis revealed a striking similarity between the score distributions of CLIP4Sketch-generated sketches and real forensic sketches. This close alignment in distribution patterns strongly indicates that our synthetic sketches achieve a high degree of realism, closely mimicking the characteristics of genuine forensic sketches.

Dataset	Description	Train/Test	No. of Images	No. of Identities
CUHK [22]	Hand drawn viewed sketches.	Train	3,600	1,800
PRIP-Composites [79]	Viewed and non-viewed sketch composites with mugshot mates. Contains software generated and hand drawn sketches.	Train	1,112	556
FS2K [78]	Artistic sketch pairs in 3 styles.	Train	4,208	2,104
TUFTS [80]	Face database for computerized sketches.	Test	224	112
IIIT-D [82]	Forensic and semi forensic pairs of sketches.	Test	144	72
WildSketch [83]	Artistic photo sketch pairs	Test	1,492	796
Our Synthetic Dataset	Mugshot and sketches in 4 styles.	Train	245,376	27,264

Table 5.1 Summary of the datasets used for training and evaluating the CLIP4Sketch model. The datasets include publicly available sketch and mugshot/photo collections as well as a large synthetic dataset generated using our proposed model.

5.3.2 Feature Space Visualization

To complement our quantitative analysis, we utilized t-SNE [70] for visualizing the similarity between generated and real sketches in the VGG latent space. The resulting t-SNE plots, as illustrated in Figure 5.3, demonstrated remarkably similar clustering behaviors between the two sets of sketches. This visualization revealed that CLIP4Sketch-generated sketches form clusters that closely resemble those of real forensic sketches in the feature space with some deviations. Such clustering patterns provide strong evidence that our generated sketches share some fundamental feature characteristics with their real counterparts. This suggests that our synthetic sketches could serve as valuable assets in enhancing sketch-to-mugshot matching performance. By generating sketches that closely emulate the characteristics of real forensic sketches, CLIP4Sketch opens up new possibilities for augmenting training datasets, improving matching algorithms, and ultimately advancing the field of forensic sketch analysis.

5.4 Sketch to Face Matching

We also investigated the model's performance in both open-set and closed-set scenarios to provide a thorough assessment of its capabilities. Throughout our experiments, we used various datasets for training and testing, including our synthetic dataset and several publicly available sketch and mugshot

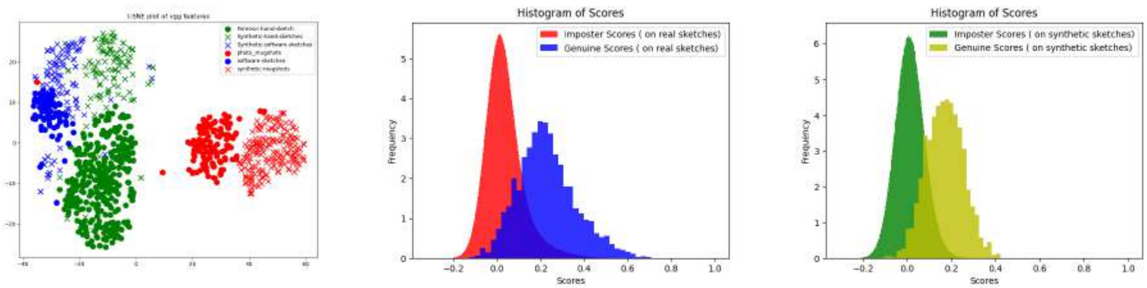

Figure 5.3 First, T-SNE plot shows modality gap in sketch images from mugshots and the similarity of our generated sketches and real sketches. The first histogram shows genuine and imposter score distributions for face-sketch pairs in real sketch datasets like CUHK [22], PRIP-Composites [79] and the second histogram shows the distribution for the dataset generated using our proposed CLIP4Sketch.

datasets. The following subsections detail our experimental setup, methodologies, and findings, providing insights into the strengths and limitations of our approach in enhancing sketch-to-mugshot matching performance.

5.4.1 Implementation Details

We conducted our experiments by finetuning pretrained AdaFace models using different datasets. We preprocess the dataset by cropping and aligning faces, resulting in 112×112 images. The three augmentations proposed in AdaFace [76] were also applied with a probability of 0.2. We use Adam optimizer [84] and a decaying learning rate strategy. It takes us 10 epochs to finetune the model with a batch size of 64. The training was done on 8 A6000 GPU's.

5.4.2 Performance Trade-offs

The experimental analysis revealed crucial insights about synthetic data's impact on sketch-to-face recognition systems. Testing different proportions of synthetic data (25%, 50%, 75%, and 100%), corresponding to 6,816 to 27,264 unique identities, demonstrated a clear trend in model performance. Increasing synthetic data showed a dual effect: it enhanced sketch-to-face matching capabilities across multiple test sets (WildSketch, IIIT-D Sketches, and TUFTS), while simultaneously causing a decline in traditional face-to-face recognition performance on the CASIAWebFace test set. This serves as evidence for fundamental tradea -off between specialized cross-modal performance and general face recognition capabilities.

5.4.3 Comparative Analysis

When evaluated against other synthetic data generation approaches, the CLIP4Sketch-generated dataset demonstrated superior performance. In comprehensive testing using a gallery of 10,000 mugshots,

Model & Dataset	TAR (%) @ FAR=0.1%			
	CASIA	WildSketch	IIIT-D Sketches	TUFTS
Pretrained Adaface	94.10	66.10	55.80	14.41
Adaface finetuned on 25% of synthetic data	53.20 ±1.5	82.70 ±2.28	44.80 ±5.60	18.44 ±3.50
Adaface finetuned on 50% of synthetic data	46.00 ±3.3	84.03 ±1.25	49.61 ±7.70	19.11 ±4.30
Adaface finetuned on 75% of synthetic data	39.40 ±1.5	86.67 ±1.83	60.5 ±5.40	19.40 ±3.68
Adaface finetuned on 100% of synthetic data	35.10	85.88	58.80	20.75

Table 5.2 Performance comparison of AdaFace model finetuned on varying amounts of synthetic data. The table shows True Accept Rate (TAR) at a False Accept Rate (FAR) of 0.1% across different datasets. The upward trend in sketch datasets shows the potential of our approach while high variance in datasets like TUFTS and IIIT-D shows the inherent challenges in them.

our model achieved an open-set FNIR of 33% and a closed-set rank-1 accuracy of 85%. This significantly outperformed both Informative Drawings (88% FNIR, 59% rank-1 accuracy) and Face-SCG (97% FNIR, 14% rank-1 accuracy).

These results demonstrate that while synthetic data is crucial for improving cross-modal recognition, its implementation requires careful consideration. The optimal solution may involve finding the right balance between synthetic and real data, depending on specific application requirements. While the current results show promising improvements over existing methods, they also indicate room for further advancement in cross-modal face recognition systems.

Model & Dataset	Closed-set Rank-1 (%)	Open-set FNIR (%) @FPIR=2%
Pretrained Adaface	72	51
Adaface finetuned on Face-Sketch-SCG [47]	14	97
Adaface finetuned on InformativeDrawings [50]	59	88
Adaface finetuned on our dataset	**85**	**33**

Table 5.3 Comparison of closed-set Rank-1 accuracy and open-set FNIR@FPIR=2% between our synthetically generated dataset and another dataset made using InformativeDrawings [50] which use GANs for sketch generation. The gallery size was 10,000, and the probe set had 980 identities.

5.5 Chapter Conclusion

The CLIP4Sketch approach addresses the key challenges of forensic sketch-to-mugshot matching. By leveraging diffusion models in conjunction with identity-preserving techniques, CLIP4Sketch is able to generate a diverse range of synthetic sketches that maintain a high degree of resemblance to real forensic sketches. Through a combination of CLIP and Adaface embeddings, the model successfully bridges the modality gap between sketches and photographs, creating a large, varied dataset with realistic stylistic differences. This synthetic dataset not only enhances data diversity but also substantially improves the performance of face recognition models in sketch-to-mugshot matching tasks, making it a valuable resource for forensic applications where annotated data is scarce.

This approach has shown that diffusion models, when combined with advanced conditioning methods, are highly effective for tasks requiring both visual fidelity and stylistic control. The quantitative results from CLIP4Sketch demonstrate considerable improvements over traditional GAN-based methods in both open-set and closed-set matching scenarios, highlighting the robustness and scalability of the model. Looking ahead, CLIP4Sketch lays the groundwork for further refinement and potential real-world application, such as in law enforcement and forensic investigation settings.

Chapter 6

Conclusion

This thesis explored the evolving domain of sketch generation and recognition, with a special focus on advancements in generative models and their applications in artistic and forensic contexts. We presented three main contributions in the field: a novel face cartoonization approach leveraging StyleGAN's latent space, the Portrait Sketching StyleGAN (PS-StyleGAN) method for stylized sketch generation, and CLIP4Sketch, a diffusion-based approach for forensic sketch generation and sketch-to-mugshot matching.

Our initial focus was on the unique challenges of face cartoonization. Through an innovative use of StyleGAN's latent space, we achieved high-quality cartoonization results that maintain identity and support pose variation. The encoder-based approach minimized the need for model fine-tuning, allowing effective adaptation of a pre-trained model for this novel task.

Building on this, we introduced PS-StyleGAN, a style transfer approach for sketch synthesis without requiring extensive model adjustments. By incorporating attention-based affine transformations, we were able to achieve artistically expressive, yet realistic, portrait sketches. This method maintained StyleGAN's latent space integrity, enabling compatibility with existing image recognition systems while allowing user control over stylistic elements.

Finally, we developed CLIP4Sketch, a diffusion-based model that generates synthetic sketches aligned with real forensic sketch attributes, greatly enhancing the diversity of data available for sketch-to-mugshot matching. The model combines CLIP and Adaface embeddings with text-based prompts to create a dataset of high-fidelity, stylistically varied sketches. This method significantly improved sketch-to-photo matching accuracy, especially in low-resource scenarios typical of forensic applications.

6.1 Future Directions

While our work achieved promising results, several areas for further exploration remain. For example, developing methods to preserve identities that not seen in training data, while still maintaining computational efficiency, could further enhance applicability of generated sketches. Additionally, applying diffusion-based models directly to real-time forensic investigations or extending them to other artistic applications presents exciting possibilities for future work. Collectively, the contributions of this thesis advance the state of sketch generation and recognition, bridging artistic and forensic needs. These methods not only broaden the capabilities of generative models in sketch synthesis but also highlight the potential of the proposed advancements in fields ranging from creative industries to forensic sciences.

List of Related Publications

1. Kushal Kumar Jain, Ankith Varun, and Anoop Namboodiri, **"PS-StyleGAN: Illustrative Portrait Sketching using Attention-Based Style Adaptation"**, in proceedings of *27th International Conference on Pattern Recognition*, Kolkata, 2024.

2. Kushal Kumar Jain, Steven Grosz, Anoop M. Namboodiri and Anil K. Jain, **"CLIP4Sketch: Enhancing Sketch to Mugshot Matching through Dataset Augmentation using Diffusion Models"**,in proceedings of *2024 IEEE International Joint Conference on Biometrics (IJCB)*, Buffalo, NY, USA, 2024.

3. Kushal Kumar Jain, Ankith Varun, and Anoop Namboodiri, **"Face Cartoonisation For Various Poses Using StyleGAN"**, preprint at arXiv, 2023.

Bibliography

[1] A. Leroi-Gourhan, “The archaeology of lascaux cave,” *Scientific American*, vol. 246, no. 6, pp. 104–113, 1982. [Online]. Available: http://www.jstor.org/stable/24966617

[2] O. BENESCH, “Leonardo da vinci and the beginning of scientific drawing,” *American Scientist*, vol. 31, no. 4, pp. 311–328, 1943. [Online]. Available: http://www.jstor.org/stable/27826007

[3] D. Y. Tsao and M. S. Livingstone, “Mechanisms of face perception.” *Annual review of neuroscience*, vol. 31, pp. 411–37, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:14760952

[4] A. Hertzmann, “The role of edges in line drawing perception,” *Perception*, vol. 50, pp. 266 – 275, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231698870

[5] B. Sayim and P. Cavanagh, “What line drawings reveal about the visual brain,” *Frontiers in Human Neuroscience*, vol. 5, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:263708652

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 674–10 685, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245335280

[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:14124313

[8] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:206593710

[9] X. Huang and S. J. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1510–1519, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:6576859

[10] S. Liu, T. Lin, D. He, F. Li, M. Wang, X. Li, Z. Sun, Q. Li, and E. Ding, "Adaattn: Revisit attention mechanism in arbitrary neural style transfer," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6629–6638, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236956663

[11] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," in *33rd International Conference on Machine Learning, ICML 2016*, 2016, pp. 2027–2041.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:261560300

[13] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *ArXiv*, vol. abs/1411.1784, 2014.

[14] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:54482423

[15] ——, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *ArXiv*, vol. abs/2006.11239, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219955663

[17] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *ArXiv*, vol. abs/2010.02502, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:222140788

[18] P. Dhariwal and A. Q. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=AAWuCvzaVt

[19] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *ArXiv*, vol. abs/2204.06125, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248097655

[20] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," *ArXiv*, vol. abs/2205.11487, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248986576

[21] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *Proceedings Ninth IEEE International Conference on Computer Vision*, 2003, pp. 687–694 vol.1.

[22] X. Wang and X. Tang, “Face photo-sketch synthesis and recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009.

[23] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, “A nonlinear approach for face sketch synthesis and recognition,” *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, pp. 1005–1010 vol. 1, 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:14224583

[24] D. DeCarlo, A. Finkelstein, S. Rusinkiewicz, and A. Santella, “Suggestive contours for conveying shape,” *ACM SIGGRAPH 2003 Papers*, 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:1485904

[25] S. Rusinkiewicz, D. DeCarlo, and A. Finkelstein, “Line drawings from 3d models,” in *International Conference on Computer Graphics and Interactive Techniques*, 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:10994464

[26] Y. Ohtake, A. G. Belyaev, and H.-P. Seidel, “Ridge-valley lines on meshes via implicit surface fitting,” *ACM SIGGRAPH 2004 Papers*, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:8500135

[27] D. Liu, M. Fisher, A. Hertzmann, and E. Kalogerakis, “Neural strokes: Stylized line drawing of 3d shapes,” *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14 184–14 193, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238531682

[28] I. Berger, A. Shamir, M. Mahler, E. J. Carter, and J. K. Hodgins, “Style and abstraction in portrait sketching,” *ACM Transactions on Graphics (TOG)*, vol. 32, pp. 1 – 12, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:17238299

[29] A. Sharma and D. W. Jacobs, “Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch,” *CVPR 2011*, pp. 593–600, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:324005

[30] W. Zhang, X. Wang, and X. Tang, “Coupled information-theoretic encoding for face photo-sketch recognition,” *CVPR 2011*, pp. 513–520, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:2031123

[31] B. Klare, Z. Li, and A. K. Jain, “Matching forensic sketches to mug shot photos,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 639–646, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:15597758

[32] H. Han, B. Klare, K. Bonnen, and A. K. Jain, “Matching composite sketches to face photos: A component-based approach,” *IEEE Transactions on Information Forensics and Security*, vol. 8, pp. 191–204, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:13547735

[33] P. Mittal, A. Jain, G. Goswami, R. Singh, and M. Vatsa, "Recognizing composite sketches with digital face images via ssd dictionary," *IEEE International Joint Conference on Biometrics*, pp. 1–6, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:17728373

[34] C. Peng, N. Wang, J. Li, and X. Gao, "Dlface: Deep local descriptor for cross-modality face recognition," *Pattern Recognit.*, vol. 90, pp. 161–171, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:126728521

[35] C. Galea and R. A. Farrugia, "Forensic face photo-sketch recognition using a deep learning-based architecture," *IEEE Signal Processing Letters*, vol. 24, pp. 1586–1590, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:23362178

[36] H. Kazemi, S. Soleymani, A. Dabouei, S. M. Iranmanesh, and N. M. Nasrabadi, "Attribute-centered loss for soft-biometrics guided face sketch-photo recognition," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 612–6128, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4710805

[37] S. Bae, N. U. Din, H. J. Park, and J. Yi, "Exploiting an intermediate latent space between photo and sketch for face photo-sketch recognition," *Sensors (Basel, Switzerland)*, vol. 22, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252617411

[38] S. Nagpal, M. Singh, R. Singh, M. Vatsa, A. Noore, and A. Majumdar, "Face sketch matching via coupled deep transform learning," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5429–5438, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:9759547

[39] J. Xu, X. Xue, Y. Wu, and X. Mao, "Matching a composite sketch to a photographed face using fused hog and deep feature models," *The Visual Computer*, vol. 37, pp. 765–776, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:224839101

[40] T. Chugh, M. Singh, S. Nagpal, R. Singh, and M. Vatsa, "Transfer learning based evolutionary algorithm for composite face sketch recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 117–125.

[41] S. Ouyang, T. M. Hospedales, Y.-Z. Song, and X. Li, "Forgetmenot: Memory-aware forensic facial sketch matching," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5571–5579, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:2826476

[42] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[43] F. Gao, Y. Zhu, C. Jiang, and N. Wang, "Human-inspired facial sketch synthesis with dynamic adaptation," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[44] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

[45] M.-Y. Liu, T. M. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Neural Information Processing Systems*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3783306

[46] J. Kim, M. Kim, H. Kang, and K. H. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *International Conference on Learning Representations*, 2019.

[47] C. Chen, W. Liu, X. Tan, and K. Wong, "Semi-supervised cycle-gan for face photo-sketch translation in the wild," in *Computer Vision and Image Understanding (CVIU)*, 2023.

[48] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 735–10 744, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:194358484

[49] D.-P. Fan, Z. Huang, P. Zheng, H. Liu, X. Qin, and L. V. Gool, "Facial-sketch synthesis: A new challenge," *Machine Intelligence Research*, vol. 19, pp. 257 – 287, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:248987735

[50] C. Chan, F. Durand, and P. Isola, "Learning to generate line drawings that convey geometry and semantics," in *CVPR*, 2022.

[51] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231591445

[52] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan: How to embed images into the stylegan latent space?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[53] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *European Conference on Computer Vision*, 2016.

[54] Z. C. Lipton and S. Tripathi, "Precise recovery of latent vectors from generative adversarial networks," *ArXiv*, vol. abs/1702.04782, 2017.

[55] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2287–2296, 2020.

[56] E. Collins, R. Bala, B. Price, and S. Süsstrunk, "Editing in style: Uncovering the local semantics of gans," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5770–5779, 2020.

[57] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, "Pastiche master: Exemplar-based high-resolution portrait style transfer," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7683–7692, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247627720

[58] M. J. Chong and D. A. Forsyth, "Jojogan: One shot face stylization," *ArXiv*, vol. abs/2112.11641, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245385527

[59] X. Hu, Q. Huang, Z. Shi, S. Li, C. Gao, L. Sun, and Q. Li, "Style transformer for image inversion and editing," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 327–11 336, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247450902

[60] H. Li, J. Liu, Y. Bai, H. Wang, and K. Mueller, "Transforming the latent space of stylegan for real face editing," *The Visual Computer*, pp. 1–16, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235254637

[61] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *ArXiv*, vol. abs/2208.01618, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251253049

[62] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22 500–22 510, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251800180

[63] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *ArXiv*, vol. abs/2308.06721, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260886966

[64] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and Y. Shan, "Photomaker: Customizing realistic human photos via stacked id embedding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[65] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[66] Y. Nitzan, A. H. Bermano, Y. Li, and D. Cohen-Or, "Disentangling in latent space by harnessing a pretrained generator," *ArXiv*, vol. abs/2005.07728, 2020.

[67] Z. Feng, J. Kittler, M. Awais, P. Huber, and X. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2235–2245, 2017.

[68] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9240–9249, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:198897678

[69] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain GAN inversion for real image editing," *CoRR*, vol. abs/2004.00049, 2020.

[70] L. van der Maaten and G. E. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:5855042

[71] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4766599

[72] D.-P. Fan, S. Zhang, Y.-H. Wu, Y. Liu, M.-M. Cheng, B. Ren, P. L. Rosin, and R. Ji, "Scoot: A perceptual metric for facial sketches," in *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[73] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, pp. 2378–2386, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:10649298

[74] J. Deng, J. Guo, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2018.

[75] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 5962–5979, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:252223915

[76] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 729–18 738, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247940176

[77] M. Kim, F. Liu, A. Jain, and X. Liu, "Dcface: Synthetic face generation with dual condition diffusion model," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR), pp. 12 715–12 725, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258170376

[78] D.-P. Fan, Z. Huang, P. Zheng, H. Liu, X. Qin, and L. V. Gool, "Facial-sketch synthesis: A new challenge," *Machine Intelligence Research*, vol. 19, pp. 257–287, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:248987735

[79] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain, "The facesketchid system: Matching facial composites to mugshots," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2248–2263, 2014.

[80] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani, and X. Yuan, "A comprehensive database for benchmarking imaging systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 3, pp. 509–520, 2020.

[81] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[82] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized mcwld for matching sketches with digital face images," *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 1522–1535, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:2147615

[83] L. Nie, L. Liu, Z. Wu, and W. Kang, "Unconstrained face sketch synthesis via perception-adaptive network and a new benchmark," *Neurocomputing*, 2022.

[84] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.